Heidelberg Institute for Theoretical Studies

HITS

Joint Entity Disambiguation and Clustering

Angela Fahrni, Thierry Göckel and Michael Strube

Heidelberg Institute for Theoretical Studies gGmbH

Heidelberg, Germany

n of NILs, Clustering, Disambiguation, Recognition of NILs, Clustering, Disambigua
, Disambiguation, Recognition of NILs, Clustering, Recognition of NILs, Clustering,
n, Recognition of NILs, Clustering, Disambiguation, Recognition of NILs, Clustering
of NILs, Clustering, Disambiguation, Recognition of NILs, Clustering, Disambiguatio
ring, Disambiguation, Recognition of NILs, Clustering, Recognition of NILs, Clusteri
biguation, Recognition of NILs, Clustering, Disambiguation, Recognition of NILs, Cl
nition of NILs, Clustering, Disambiguation, Recognition of NILs, Clustering, Disamb
ring, Disambiguation, Recognition of NILs, Clustering, Recognition of NILs, Clusteri
uation, Recognition of NILs, Clustering, Disambiguation, Recognition of NILs, Cluste
n of NILs, Clustering, Disambiguation, Recognition of NILs, Clustering, Disambigua
, Disambiguation, Recognition of NILs, Clustering, Recognition of NILs, Clustering,
n, Recognition of NILs, Clustering, Disambiguation, Recognition of NILs, Clustering
of NILs, Clustering, Disambiguation, Recognition of NILs, Clustering, Disambiguatio
ring, Disambiguation, Recognition of NILs, Clustering, Recognition of NILs, Clusteri
biguation, Recognition of NILs, Clustering, Disambiguation, Recognition of NILs, Cl
nition of NILs, Clustering, Disambiguation, Recognition of NILs, Clustering, Disamb
ring, Disambiguation, Recognition of NILs, Clustering, Recognition of NILs, Clusteri
uation, Recognition of NILs, Clustering, Disambiguation, Recognition of NILs, Cluste
n of NILs, Clustering, Disambiguation, Recognition of NILs, Clustering, Disambigua
, Disambiguation, Recognition of NILs, Clustering, Recognition of NILs, Clustering,

Common and Proper Nouns

Common and Proper Nouns

Focus on Modelling

Common and Proper Nouns

Focus on Modelling

Joint Approach

Common and Proper Nouns

Focus on Modelling

Joint Approach

Markov Logic

Common and Proper Nouns

Focus on Modelling

Joint Approach

Low Average Ambiguity

Markov Logic

Common and Proper Nouns

Focus on Modelling

Joint Approach

Low Average Ambiguity

Markov Logic

Training on 500 English Wikipedia Articles

Recognition of NILs, Clustering, Disambiguation

Common and Proper Nouns

Focus on Modelling

Joint Approach

Low Average Ambiguity

Markov Logic

Competitive Results

Training on 500 English Wikipedia Articles

Common and Proper Nouns

Focus on Modelling

Joint Approach

Low Average Ambiguity

Markov Logic

Competitive Results

Training on 500 English Wikipedia Articles

Recognition of NILs, Clustering, Disambiguation

Text 1

Within the States, American crocodiles live in Florida.

Recently, the biologist Aldecoa captured an older crocodile in the sunshine state.

Text 2

The biologist Aldecoa caught the hatchlings.

Text 1

Within the $\boxed{\text{States}}$, $\boxed{\text{American crocodiles}}$ live in $\boxed{\text{Florida}}$.

Recently, the $\boxed{\text{biologist}}$ $\boxed{\text{Aldecoa}}$ captured an older $\boxed{\text{crocodile}}$ in the $\boxed{\text{sunshine state}}$.

Text 2

The $\boxed{\text{biologist}}$ $\boxed{\text{Aldecoa}}$ caught the $\boxed{\text{hatchlings}}$.

Knowledge Base

- United States
- State (Polity)
- State of matter
- Florida (US State)
- Florida (Puerto Rico)
- American Crocodiles (Animal)
- Crocodile (Locomotive)
- René Lacoste (Tennis player)
- Biologist
- Hatchling
- Ignacio Aldecoa (Spanish Author)
- Emilio Aldecoa (Football player)

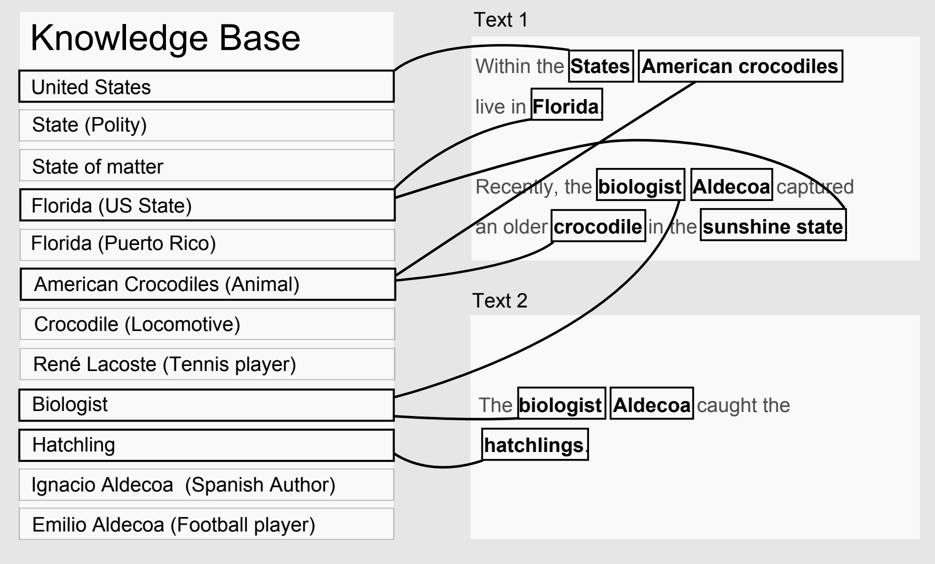
Text 1

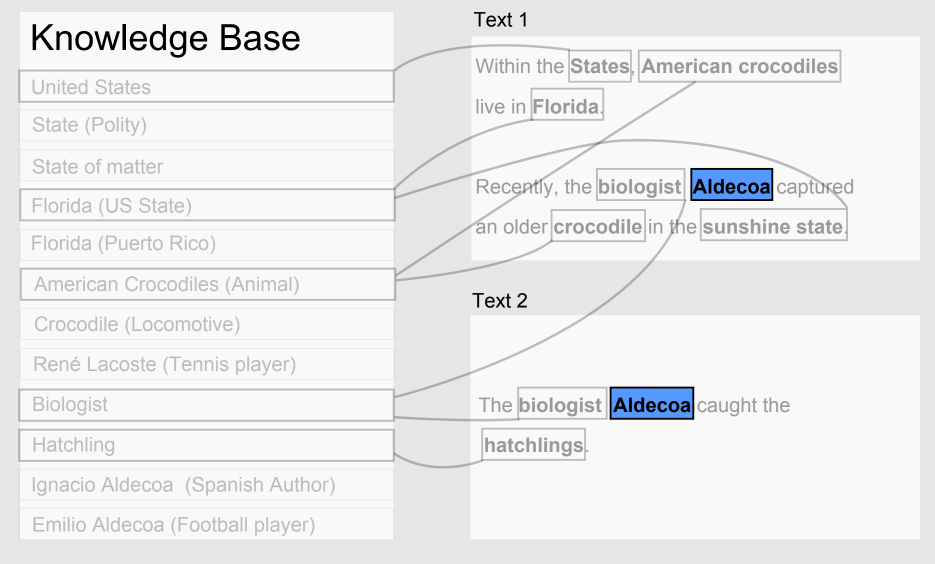
Within the **States**, **American crocodiles** live in **Florida**.

Recently, the **biologist** **Aldecoa** captured an older **crocodile** in the **sunshine state**.

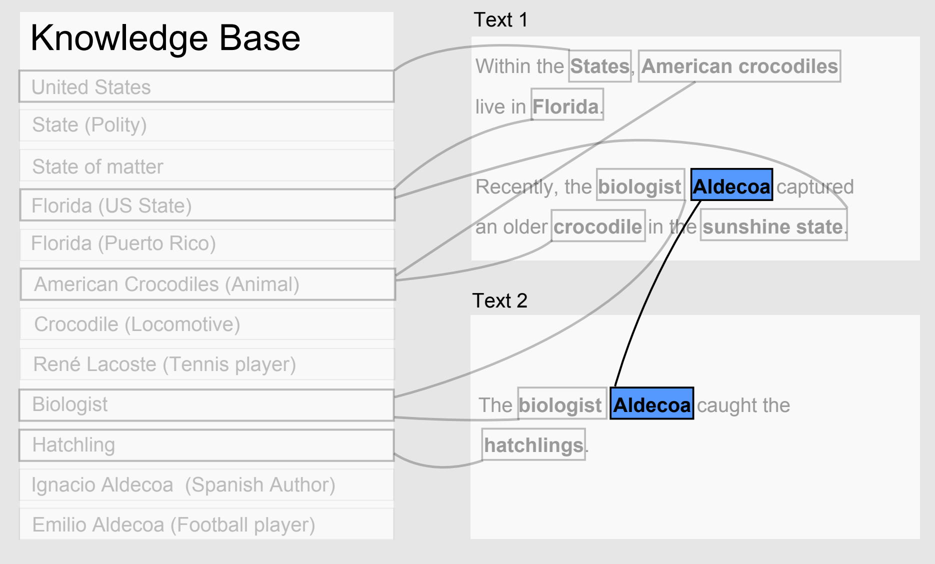
Text 2

The **biologist** **Aldecoa** caught the **hatchlings**.


Knowledge Base
United States
State (Polity)
State of matter
Florida (US State)
Florida (Puerto Rico)
American Crocodiles (Animal)
Crocodile (Locomotive)
René Lacoste (Tennis player)
Biologist
Hatchling
Ignacio Aldecoa (Spanish Author)
Emilio Aldecoa (Football player)
Text 1
Within the States, American crocodiles live in Florida.
Recently, the biologist Aldecoa captured an older crocodile in the sunshine state.
Text 2
The biologist Aldecoa caught the hatchlings.



Knowledge Base
United States
State (Polity)
State of matter
Florida (US State)
Florida (Puerto Rico)
American Crocodiles (Animal)
Crocodile (Locomotive)
René Lacoste (Tennis player)
Biologist
Hatchling
Ignacio Aldecoa (Spanish Author)
Emilio Aldecoa (Football player)
Text 1
Within the States, American crocodiles live in Florida.
Recently, the biologist Aldecoa captured an older crocodile in the sunshine state.
Text 2
The biologist Aldecoa caught the hatchlings.


Knowledge Base
United States
State (Polity)
State of matter
Florida (US State)
Florida (Puerto Rico)
American Crocodiles (Animal)
Crocodile (Locomotive)
René Lacoste (Tennis player)
Biologist
Hatchling
Ignacio Aldecoa (Spanish Author)
Emilio Aldecoa (Football player)
Text 1
Within the States, American crocodiles
live in Florida.
Recently, the biologist Aldecoa captured
an older crocodile in the sunshine state.
Text 2
The biologist Aldecoa caught the
hatchlings.

Common and Proper Nouns

Focus on Modelling

Joint Approach

Low Average Ambiguity

Markov Logic

Competitive Results

Training on 500 English Wikipedia Articles

Our Last Year's Approach

Cascaded Approach

for each Text *t*

for all Noun n in Text *t*

Entity candidates identification

end for

↓

NIL detection

↓

Entity disambiguation

end for

↓

Clustering of NILs

Our Last Year's Approach

Cascaded Approach

for each Text *t*

for all Noun n in Text *t*

Entity candidates identification

end for

↓

NIL detection

↓

Entity disambiguation

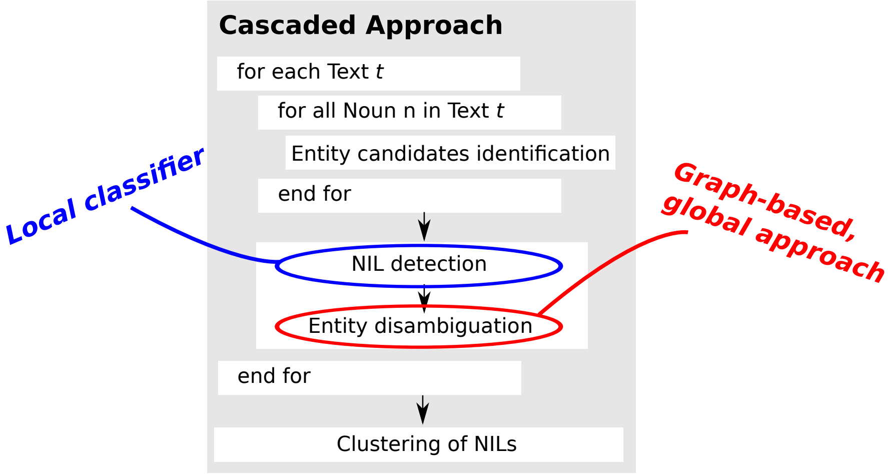
end for

↓

Clustering of NILs

Local classifier

Our Last Year's Approach



Our Last Year’s Approach

Cascaded Approach

for each Text *t*

for all Noun n in Text *t*

Entity candidates identification

end for

NIL detection

Entity disambiguation

end for

Clustering of NILs

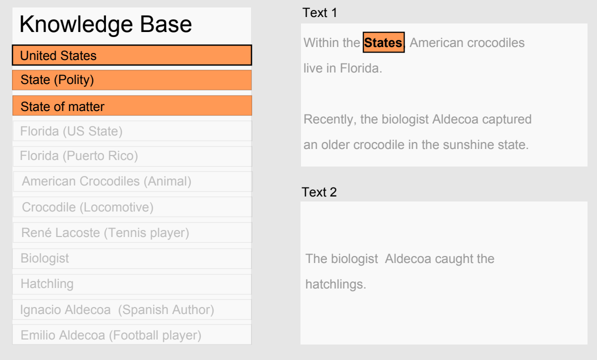
Local classifier

Graph-based, global approach

Graph-based clustering

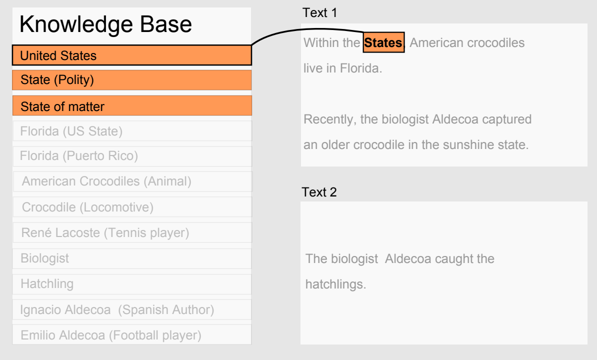
Concept/Entity Disambiguation

Local approaches: Local supervised classification or ranking approaches



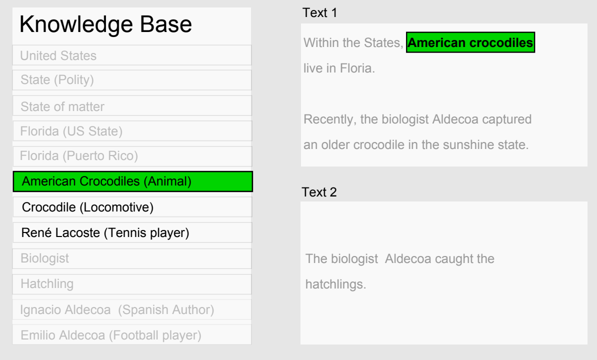
Concept/Entity Disambiguation

Local approaches: Local supervised classification or ranking approaches



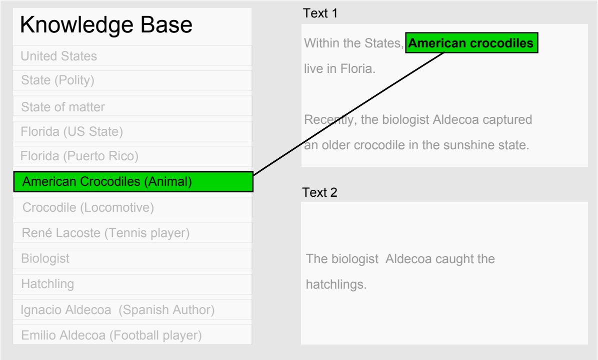
Concept/Entity Disambiguation

Local approaches: Local supervised classification or ranking approaches



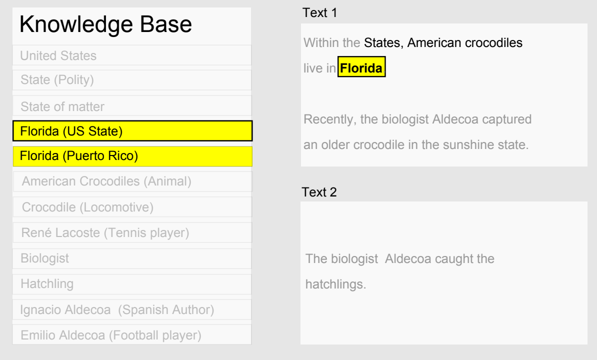
Concept/Entity Disambiguation

Local approaches: Local supervised classification or ranking approaches



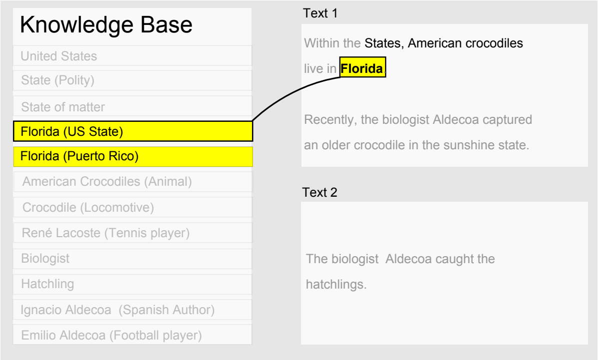
Concept/Entity Disambiguation

Local approaches: Local supervised classification or ranking approaches



Concept/Entity Disambiguation

Local approaches: Local supervised classification or ranking approaches



Concept/Entity Disambiguation

Global approaches: Collective classification approaches

Knowledge Base

- United States
- State (Polity)
- State of matter
- Florida (US State)
- Florida (Puerto Rico)
- American Crocodiles (Animal)
- Crocodile (Locomotive)
- René Lacoste (Tennis player)
- Biologist
- Hatchling
- Ignacio Aldecoa (Spanish Author)
- Emilio Aldecoa (Football player)

Text 1

Within the **States** **American crocodiles** live in **Florida**.

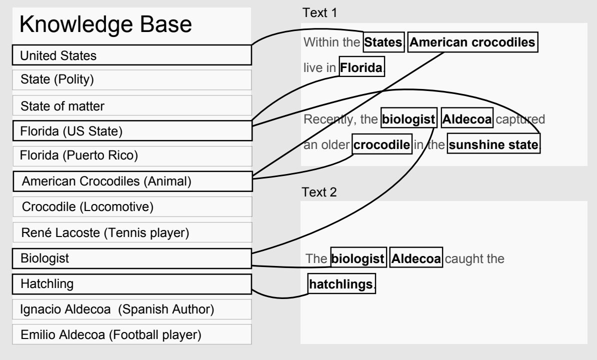
Recently, the **biologist** **Aldecoa** captured an older **crocodile** in the **sunshine state**

Text 2

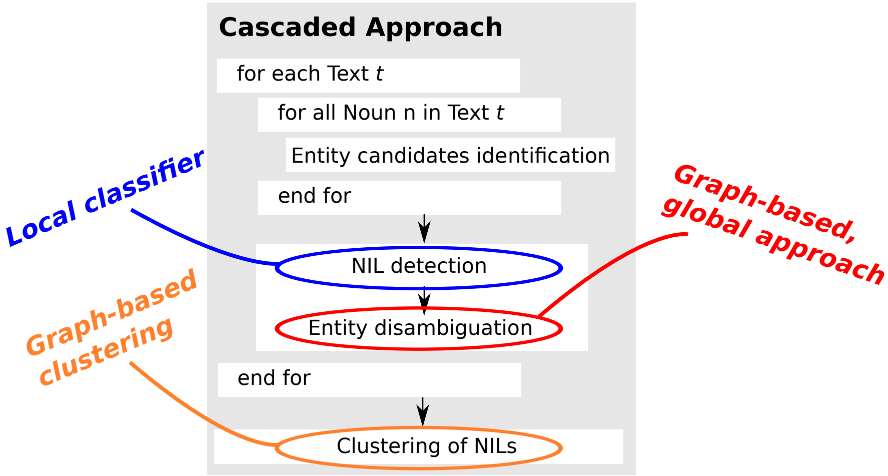
The **biologist** **Aldecoa** caught the **hatchlings**.

Concept/Entity Disambiguation

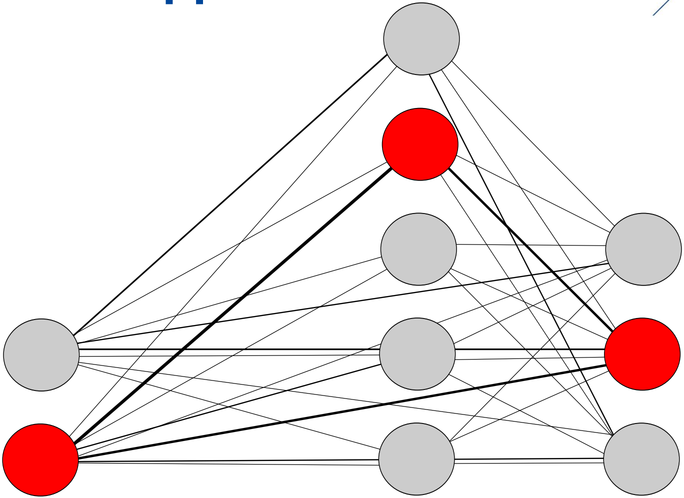
Global approaches: Collective classification approaches



Our Last Year's Approach



Graph-based Approach



Recently, the **[**biologist**]** **[**Aldecoa**]** captured a **[**crocodile**]** in the **[**sunshine state**]**

Summary Cascaded Approach

- Global approach with competitive results (best system at NTCIR 9, between median and best system in the Chinese cross-lingual entity linking task at TAC 2011)
- Error propagation
- Training: how to integrate more (local) features?
- Non-pairwise features?

Error Propagation

Cascaded Approach

for each Text *t*

for each Noun n in Text *t*

Entity candidates identification — **97%**

↓

NIL detection

↓

Entity disambiguation

end for

end for

↓

Clustering of NILs

Error Propagation

Cascaded Approach

for each Text *t*

for each Noun n in Text *t*

Entity candidates identification — **97%**

↓

NIL detection — **80%**

↓

Entity disambiguation

end for

end for

↓

Clustering of NILs

Error Propagation

Cascaded Approach

for each Text *t*

for each Noun n in Text *t*

Entity candidates identification 97%

↓

NIL detection 80%

↓

Entity disambiguation 72%

end for

end for

↓

Clustering of NILs

Error Propagation

Cascaded Approach

for each Text *t*

for each Noun n in Text *t*

Entity candidates identification 97%

↓

NIL detection 80%

↓

Entity disambiguation 72%

56%

end for

end for

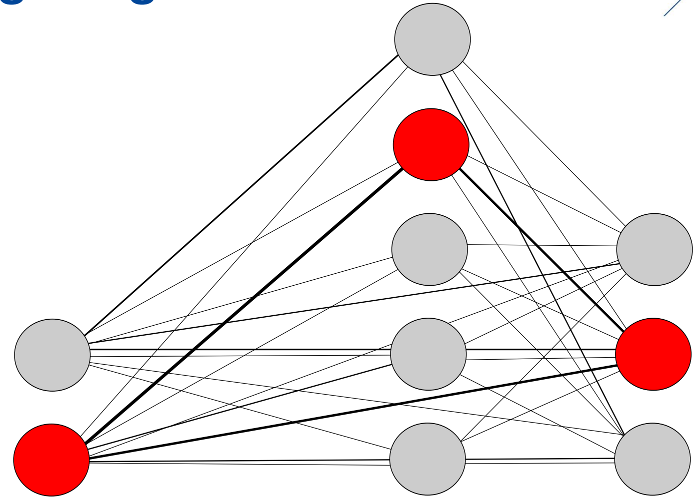
↓

Clustering of NILs

Summary Cascaded Approach

- Global approach with competitive results (best system at NTCIR 9, between median and best system in the Chinese cross-lingual entity linking task at TAC 2011)
- Error propagation
- Training: how to integrate more (local) features?
- Non-pairwise features?

Learning Weights

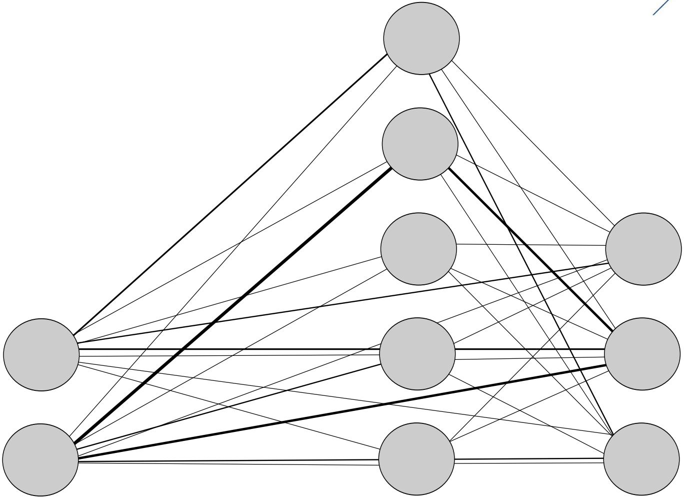


Recently, the **[**biologist**]** **[**Aldecoa**]** captured a **[**crocodile**]** in the **[**sunshine state**]**

Summary Cascaded Approach

- Global approach with competitive results (best system at NTCIR 9, between median and best system in the Chinese cross-lingual entity linking task at TAC 2011)
- Error propagation
- Training: how to integrate more (local) features?
- Non-pairwise features?

Non-pairwise Features?



Recently, the **[**biologist**]** **[**Aldecoa**]** captured a **[**crocodile**]** in the **[**sunshine state**]**

Summary Cascaded Approach

- Global approach with competitive results (best system at NTCIR 9, between median and best system in the Chinese cross-lingual entity linking task at TAC 2011)
- Error propagation
- Training: how to integrate more (local) features?
- Non-pairwise features?

Common and Proper Nouns

Focus on Modelling

Joint Approach

Low Average Ambiguity

Markov Logic

Competitive Results

Training on 500 English Wikipedia Articles

Novel Approach

Cascaded Approach

for each Text *t*

for each Noun n in Text *t*

Entity candidates identification
↓
NIL detection
↓
Entity disambiguation

end for

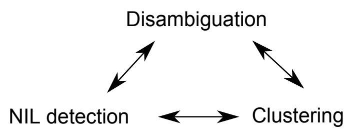
end for
↓
Clustering of NILs

Joint Approach

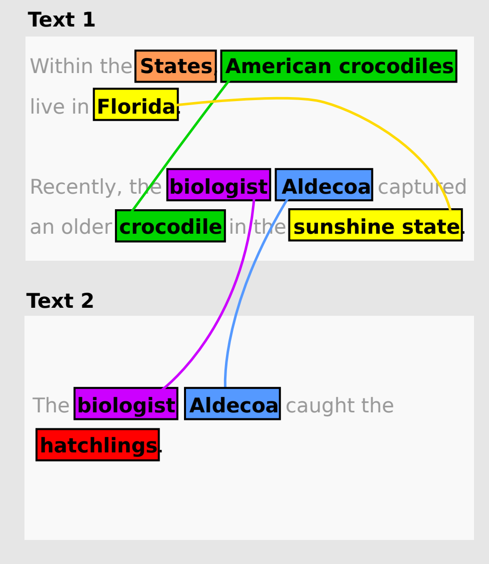
for each Noun n in each Text

Entity candidates identification

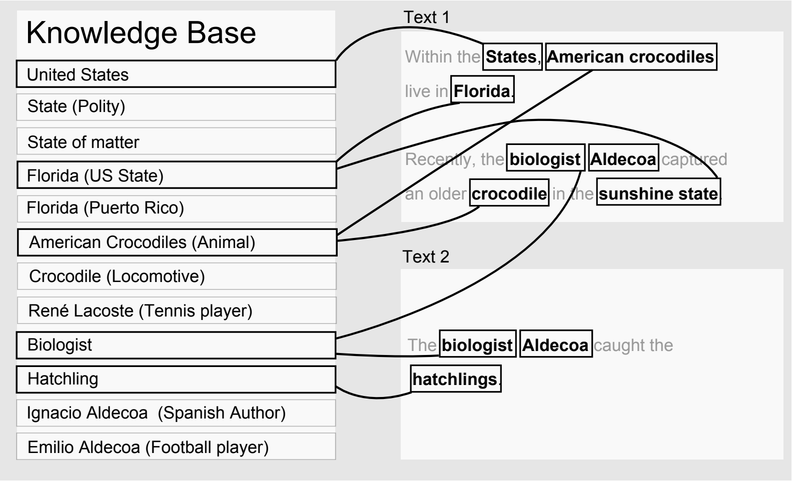
end for
↓



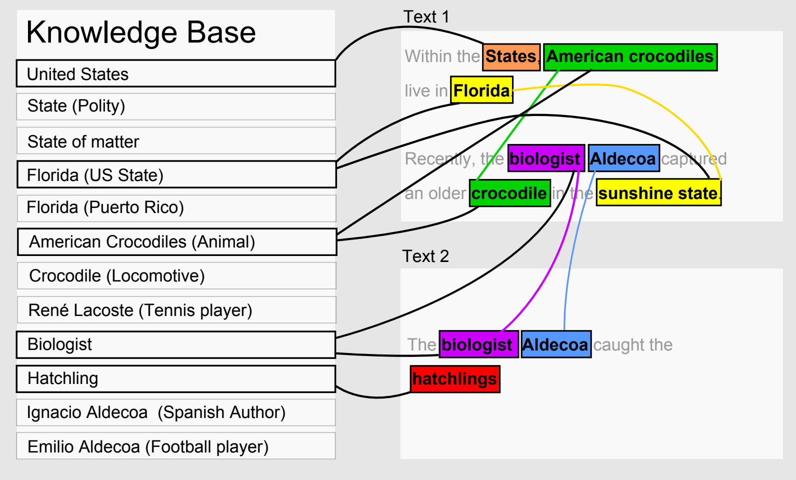
Joint Approach



Joint Approach



Joint Approach



Joint Approach

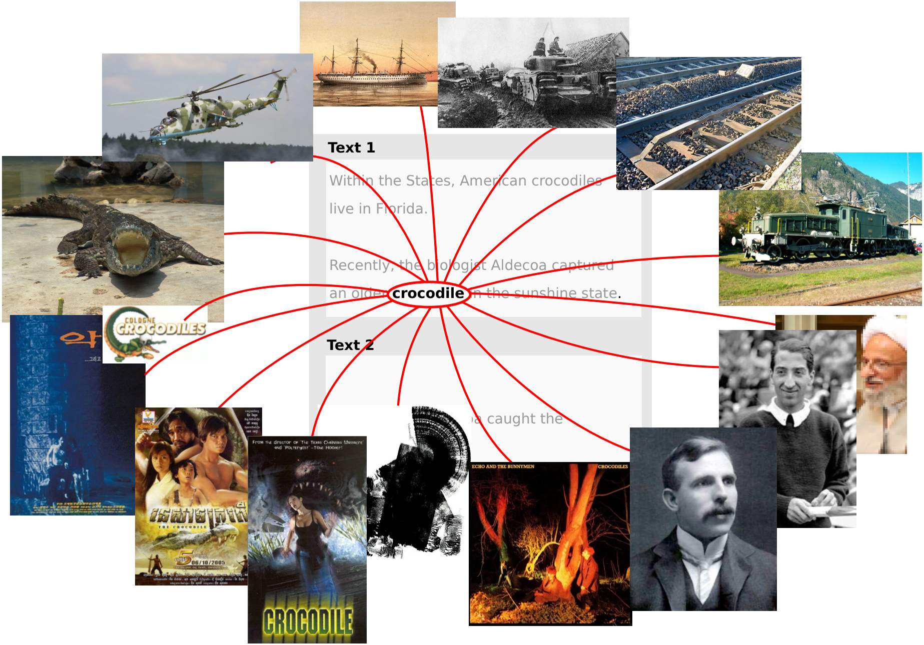
Text 1

Within the States, American crocodiles live in Florida.

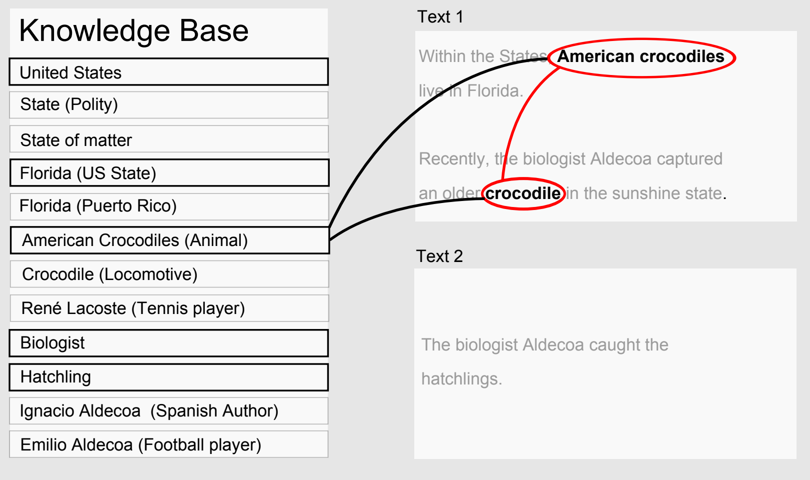
Recently, the biologist Aldecoa captured an older **crocodile** in the sunshine state.

Text 2

The biologist Aldecoa caught the hatchlings.


Text 1
Within the States, American crocodiles live in Florida.
Recently, the biologist Aldecoa captured an older crocodile in the sunshine state.
Text 2
a caught the
COLOGNE CROCODILES
THE CROCODILE
06/10/2005
From the director of 'The Texas Chainsaw Massacre' and 'Poltergeist' – Tobe Hooper
CROCODILE
ECHO AND THE BUNNYMEN
CROCODILES

Joint Approach



Requirements

- Joint inference
  - disambiguation, recognition of NILs and clustering
- Non-pairwise features

Requirements

- Joint inference
  - disambiguation, recognition of NILs and clustering
- Non-pairwise features

$\rightarrow$ **Markov Logic**

Common and Proper Nouns

Focus on Modelling

Joint Approach

Low Average Ambiguity

Markov Logic

Competitive Results

Training on 500 English Wikipedia Articles

Markov Logic

- Markov Logic (ML) combines **first-order logic** with **probabilities**

- A Markov Logic Network (MLN) is a set of pairs (F_i, w_i)
  - F_i: first-order formula
  - w_i: weight $w_i \in \mathbb{R}$ associated with formula F_i

Domingos and Lowd 2009

Markov Logic Networks (MLN)

A MLN is a **template** for constructing a Markov network. Given a set of constants C a Markov network can be defined in the following way:

- Binary vertex for each possible grounding of each predicate:
  - If ground predicate is *true* $\rightarrow 1$
  - If ground predicate is *false* $\rightarrow 0$
- One feature for each possible grounding of each formula F_i:
  - If ground formula is *true* $\rightarrow 1$
  - If ground predicate is *false* $\rightarrow 0$
  - Weight: w_i

Probability Distribution

Probability for a possible world x specified by the ground Markov network:

$$P(X = x) = \frac{1}{Z} \exp\left(\sum_i w_i n_i(x)\right)$$

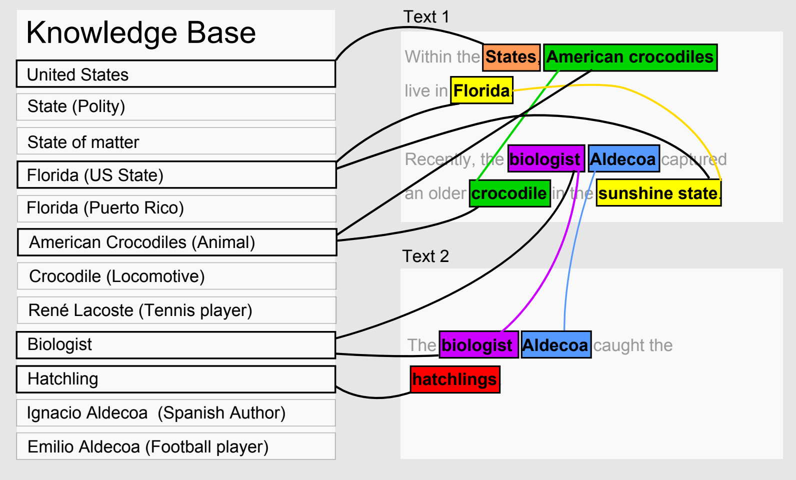
$n_i(x)$ is the number of true groundings of F_i in x
w_i is the weight of F_i

Partition function Z:

$$Z = \sum_{x \in X} \exp\left(\sum_i w_i n_i(x)\right)$$

Domingos and Lowd 2009

Disambiguation and Clustering using Markov Logic



Disambiguation and Clustering using Markov Logic

Hidden Predicates

$hasEntity(m, e)$

$hasSameEntity(m, n)$

Disambiguation and Clustering using Markov Logic

Hidden Predicates

$hasEntity(m, e)$

$hasSameEntity(m, n)$

Hard Constraints

At most one entity/concept per mention:
$\forall m \in M: \ |\{e \in E : hasEntity(m, e)\}| \leq 1$

Symmetry:
$\forall m, n \in M: \ m \neq n \wedge hasSameEntity(m, n) \rightarrow hasSameEntity(n, m)$

Transitivity:
$\forall m, n, l \in M: \ m \neq n \wedge m \neq l \wedge n \neq l$
$\wedge\, hasSameEntity(m, n) \wedge hasSameEntity(n, l) \rightarrow hasSameEntity(m, l)$

Disambiguation and Clustering using Markov Logic

Hard Constraints (continued)

All members of a cluster refer to the same entity/concept:

$\forall m, n \in M: \; m \neq n \wedge hasSameEntity(m,n) \wedge hasEntity(m,e)$
$\rightarrow hasEntity(n,e)$

If two mentions refer to the same concept, they belong to the same cluster:

$\forall m, n \in M: \; m \neq n \wedge m \neq n \wedge hasEntity(m,e) \wedge hasEntity(n,e)$
$\rightarrow hasSameEntity(m,n)$

Formulas with Learned Weight

Local Features:

Prior probability of an entity/concept given a mention:

$\forall m \in M \, \forall e \in E_m : \; hasCommonness(m, e, s) \rightarrow hasEntity(m, e)$

Weight: $+(s \cdot w)$

Example:
crocodile: American crocodile: $s = 0.4$, Crocodile (Locomotive): $s = 0.2$, Rene Lacoste: $s = 0.1$, ...

Formulas with Learned Weight

Local Features:

Not related to context:

$\forall m \in M \, \forall e \in E_m : \; hasRelatedness(m, e, s) \wedge s = 0 \rightarrow hasEntity(m, e)$

Weight: $-(w)$

Other local features: relatedness, local context similarity, string distance.

Formulas with Learned Weight

Global Features:

Same head and m is substring of n:

$$\forall m, n \in M \, \forall e \in E_m : \; m \neq n \wedge isSubStringHeadMatch(m, n, s)$$
$$\rightarrow hasEntity(m, e) \wedge hasEntity(n, e)$$

Weight: $+(s \cdot w)$

Example:
American crocodile – crocodile; Jimmy Allen – Allen

Formulas with Learned Weight

Global Features:

Two mentions with the same string tend to refer to the same entity/concept:

$$\forall m, n \in M : \; m \neq n \wedge hasSameString(m, n, s) \rightarrow hasSameEntity(m, n)$$

Weight: $+(s \cdot w)$

Example:
crocodile – crocodile

Formulas with Learned Weight

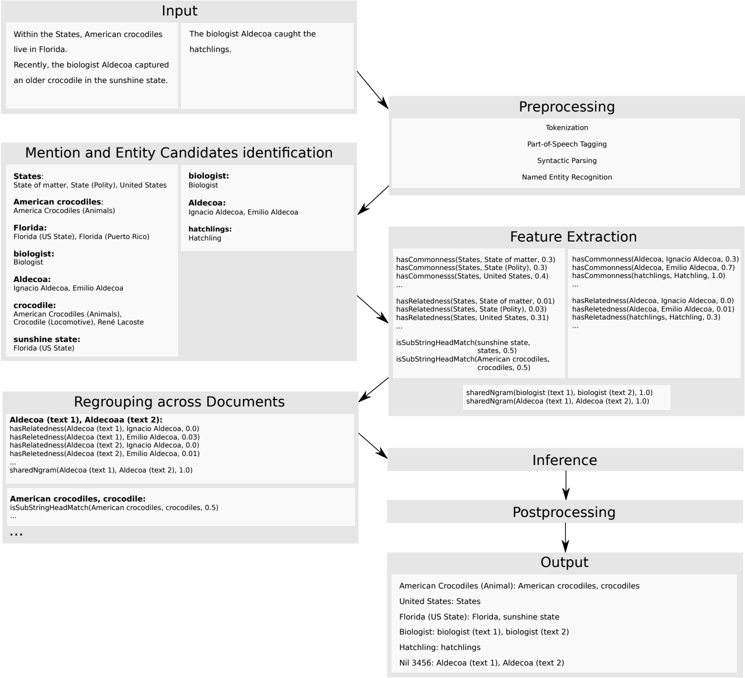
Global Features:

Two mentions in two different documents are part of the same n-gram:

$$\forall m, n \in M : \; m \neq n \wedge shareNgram(m, n, s) \rightarrow hasSameEntity(m, n)$$

Weight: $+(s \cdot w)$

Example:
biologist Aldecao – biologist Aldecao


Input
Within the States, American crocodiles live in Florida.
Recently, the biologist Aldecoa captured an older crocodile in the sunshine state.
The biologist Aldecoa caught the hatchlings.
Preprocessing
Tokenization
Part-of-Speech Tagging
Syntactic Parsing
Named Entity Recognition
Mention and Entity Candidates identification
States:
State of matter, State (Polity), United States
American crocodiles:
America Crocodiles (Animals)
Florida:
Florida (US State), Florida (Puerto Rico)
biologist:
Biologist
Aldecoa:
Ignacio Aldecoa, Emilio Aldecoa
crocodile:
American Crocodiles (Animals),
Crocodile (Locomotive), René Lacoste
sunshine state:
Florida (US State)
biologist:
Biologist
Aldecoa:
Ignacio Aldecoa, Emilio Aldecoa
hatchlings:
Hatchling
Feature Extraction
hasCommonness(States, State of matter, 0.3)
hasCommonness(States, State (Polity), 0.3)
hasCommonesss(States, United States, 0.4)
...
hasRelatedness(States, State of matter, 0.01)
hasRelatedness(States, State (Polity), 0.03)
hasRelatedness(States, United States, 0.31)
...
isSubStringHeadMatch(sunshine state, states, 0.5)
isSubStringHeadMatch(American crocodiles, crocodiles, 0.5)
hasCommonness(Aldecoa, Ignacio Aldecoa, 0.3)
hasCommonness(Aldecoa, Emilio Aldecoa, 0.7)
hasCommonness(hatchlings, Hatchling, 1.0)
...
hasRelatedness(Aldecoa, Ignacio Aldecoa, 0.0)
hasReletedness(Aldecoa, Emilio Aldecoa, 0.01)
hasReletadness(hatchlings, Hatchling, 0.3)
...
sharedNgram(biologist (text 1), biologist (text 2), 1.0)
sharedNgram(Aldecoa (text 1), Aldecoa (text 2), 1.0)
Regrouping across Documents
Aldecoa (text 1), Aldecoaa (text 2):
hasRelatedness(Aldecoa (text 1), Ignacio Aldecoa, 0.0)
hasReletedness(Aldecoa (text 1), Emilio Aldecoa, 0.03)
hasRelatedness(Aldecoa (text 2), Ignacio Aldecoa, 0.0)
hasReletedness(Aldecoa (text 2), Emilio Aldecoa, 0.01)
...
sharedNgram(Aldecoa (text 1), Aldecoa (text 2), 1.0)
American crocodiles, crocodile:
isSubStringHeadMatch(American crocodiles, crocodiles, 0.5)
...
...
Inference
Postprocessing
Output
American Crocodiles (Animal): American crocodiles, crocodiles
United States: States
Florida (US State): Florida, sunshine state
Biologist: biologist (text 1), biologist (text 2)
Hatchling: hatchlings
Nil 3456: Aldecoa (text 1), Aldecoa (text 2)

Learning and Inference

Learning:
Online training using a perceptron

Inference:
MAP inference using Cutting Planes combined with Integer Linear Programming (*Gurobi*)

Tool:
TheBeast: `http://code.google.com/p/thebeast/`

Riedel 2009

Common and Proper Nouns

Focus on Modelling

Joint Approach

Low Average Ambiguity

Markov Logic

Competitive Results

Training on 500 English Wikipedia Articles

Training and Development Data

English Wikipedia articles (featured)

After the risks caused by the flammability of hydrogen became apparent, it was replaced with helium in blimps and balloons.

```
After the risks caused by the
flammability of [[hydrogen]]
became apparent, it was replaced
with helium in [[non-rigid
airship|blimps]] and [[gas
balloon|balloons]].
```

Dataset	Documents	Mentions	in KB	NILs	Ave. Amb.
WP Training	500	46,810	43,547	3,263	2.18
WP Dev	100	7,197	6,610	587	2.11

Common and Proper Nouns

Focus on Modelling

Joint Approach

Low Average Ambiguity

Markov Logic

Competitive Results

Training on 500 English Wikipedia Articles

Going Cross-lingual

- **Mapping** of Chinese and Spanish Wikipedia articles to the English ones using interlanguage links and triangulation.

Mapped Chinese articles	48.4%
Mapped Spanish articles	60.2%

- **Disambiguation** with respect to the **mapped index**
- **English link structure** can be used to calculate relatedness
- **Trained** on the internal hyperlinks of **500 English Wikipedia articles** for all languages

Common and Proper Nouns

Focus on Modelling

Joint Approach

Low Average Ambiguity

Markov Logic

Competitive Results

Training on 500 English Wikipedia Articles

English Entity Linking Task at TAC 2012

Run	Acc	B^3 P	B^3 R	B^3 F1	B^{3+} P	B^{3+} R	B^{3+} F1
Best							0.730
Median							0.536
HITS	0.718	0.751	0.932	0.832	0.572	0.678	0.621

Chinese Entity Linking Task at TAC 2012

Run	**Micr.**	**B^3 P**	**B^3 R**	**B^3 F1**	**B^{3+} P**	**B^{3+} R**	**B^{3+} F1**
Best							0.740
HITS	0.843	0.863	0.811	0.836	0.738	0.742	0.740
EN	0.895	0.952	0.787	0.861	0.861	0.743	0.798
ZH	0.818	0.848	0.798	0.822	0.701	0.719	0.710

Spanish Entity Linking Task at TAC 2012

Run	**Micr.**	**B^3 P**	**B^3 R**	**B^3 F1**	**B^{3+} P**	**B^{3+} R**	**B^{3+} F1**
Best							0.641
HITS	0.707	0.648	0.880	0.746	0.464	0.638	0.538
HITS*	0.707	0.904	0.830	0.866	0.660	0.612	0.635

Evaluation by Different Categories (Micro-average)

CAT	EN	EN/ZH	EN/ES
PER	0.832	0.767	0.880
ORG	0.771	0.869	0.798
GPE	0.553	0.892	0.515
KB	0.598	0.794	0.519
NILs	0.853	0.912	0.859
NW	0.750	0.828	
Web	0.657	0.870	

Results on TAC 2011

	Micr.	B^3 **F**
Upperbound I	87.5	87.4
Upperbound II	97.6	97.5
Best		84.6
Median		71.6
ML Dis.	76.8	74.3
ML Dis. + NILs	78.3	75.4
ML Dis. + NILs + Clust.	82.9	80.1

Common and Proper Nouns

Focus on Modelling

Joint Approach

Low Average Ambiguity

Markov Logic

Competitive Results

Training on 500 English Wikipedia Articles

Current and future work

- Integration of more linguistic features
- More accurate relatedness measures
- Scalability

Thank you!

NLP group HITS

Mathias Niepert, University of Mannheim

Come to our poster!

Angela Fahrni
angela.fahrni@h-its.org